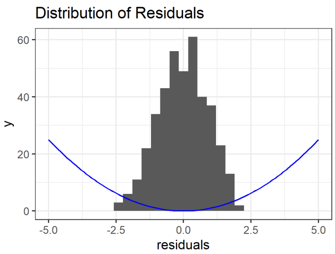
Introduction to Machine Learning

Advanced Risk Minimization
Maximum Likelihood Estimation vs. Empirical Risk Minimization



Learning goals

- Connection between maximum likelihood and risk minimization
- Correspondence between Gaussian errors and L2 loss, Laplace errors and L1 loss, and Bernoulli targets and Bernoulli/log loss

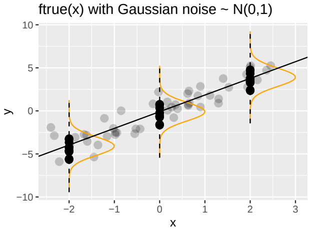
MAXIMUM LIKELIHOOD

Let's consider regression from a maximum likelihood perspective. Assume:

$$y \mid \mathbf{x} \sim p(y \mid \mathbf{x}, \boldsymbol{\theta})$$

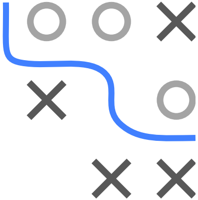
Common case: true underlying relationship f_{true} with additive noise:

$$y = f_{\text{true}}(\mathbf{x}) + \epsilon$$



where f_{true} has params $\boldsymbol{\theta}$ and ϵ a RV that follows some distribution $\mathbb{P}_\epsilon$, with $\mathbb{E}[\epsilon] = 0$. Also, assume $\epsilon \perp\!\!\!\perp \mathbf{x}$.

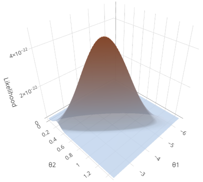
MAXIMUM LIKELIHOOD



From a statistics / maximum-likelihood perspective, we assume (or we pretend) we know the underlying distribution family $p(y \mid \mathbf{x}, \boldsymbol{\theta})$.

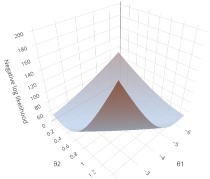
- Given i.i.d data $\mathcal{D} = \left(\left(\mathbf{x}^{(1)}, y^{(1)}\right), \ldots, \left(\mathbf{x}^{(n)}, y^{(n)}\right)\right)$ from $\mathbb{P}_{xy}$ the maximum-likelihood principle is to maximize the **likelihood**

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} p\left(y^{(i)} \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}\right)$$

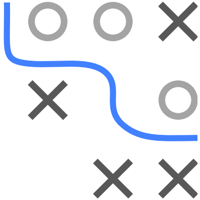


or equivalently minimize the **negative log-likelihood (NLL)**

$$-\ell(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \log p\left(y^{(i)} \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}\right)$$



MAXIMUM LIKELIHOOD



From an ML perspective we assume our hypothesis space corresponds to the space of the (parameterized) f_{true}.

- Simply define neg. log-likelihood as **loss function**

$$L\left(y, f(\mathbf{x} \mid \boldsymbol{\theta})\right) := -\log p(y \mid \mathbf{x}, \boldsymbol{\theta})$$

- Then, maximum-likelihood = ERM

$$\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} L\left(y^{(i)}, f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)$$

- NB: When we are only interested in the minimizer, we can ignore multiplicative or additive constants.
- We use $\propto$ as “proportional up to multiplicative and additive constants”

GAUSSIAN ERRORS - L2-LOSS

Assume $y = f_{\text{true}}(\mathbf{x}) + \epsilon$ with additive Gaussian errors, i.e. $\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$. Then $y \mid \mathbf{x} \sim N\left(f_{\text{true}}(\mathbf{x}), \sigma^2\right)$. The likelihood is then

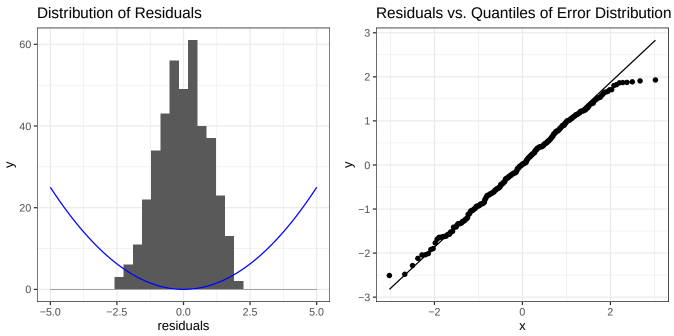
$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \prod_{i=1}^{n} p\left(y^{(i)} \,\middle|\, f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right), \sigma^2\right) \\
&\propto \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}\left(y^{(i)} - f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)^2\right)
\end{aligned}
$$

Easy to see: minimizing Gaussian NLL s is ERM with $L2$-loss:

$$
\begin{aligned}
-\ell(\boldsymbol{\theta}) &= -\log\left(\mathcal{L}(\boldsymbol{\theta})\right) \\
&\propto -\log\left(\prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}\left(y^{(i)} - f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)^2\right)\right) \\
&\propto \sum_{i=1}^{n}\left(y^{(i)} - f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)^2
\end{aligned}
$$

GAUSSIAN ERRORS - L2-LOSS

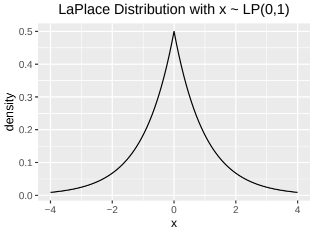
- We simulate data $y \mid \mathbf{x} \sim \mathcal{N}\left(f_{\text{true}}(\mathbf{x}), 1\right)$ with $f_{\text{true}} = 0.2 \cdot \mathbf{x}$
- Let's plot empirical errors as histogram, after fitting our model with $L2$-loss
- Q-Q-plot compares empirical residuals vs. theoretical quantiles of Gaussian

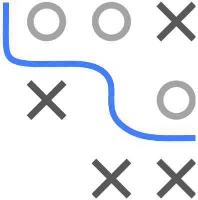


LAPLACE ERRORS - L1-LOSS

Let's consider Laplacian errors ϵ now, with density:

$$\frac{1}{2\sigma} \exp\left(-\frac{|\epsilon|}{\sigma}\right), \sigma > 0.$$





Then

$$y = f_{\text{true}}(\mathbf{x}) + \epsilon$$

also follows Laplace distribution with mean $f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)$ and scale σ.

LAPLACE ERRORS - L1-LOSS

The likelihood is then

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= \prod_{i=1}^{n} p\left(y^{(i)} \,\middle|\, f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right), \sigma\right) \\ &\propto \exp\left(-\frac{1}{\sigma}\sum_{i=1}^{n}\left|y^{(i)} - f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right|\right).\end{aligned}$$

The negative log-likelihood is

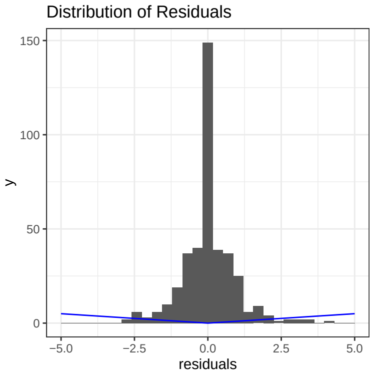
$$-\ell(\boldsymbol{\theta}) \propto \sum_{i=1}^{n}\left|y^{(i)} - f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right|.$$

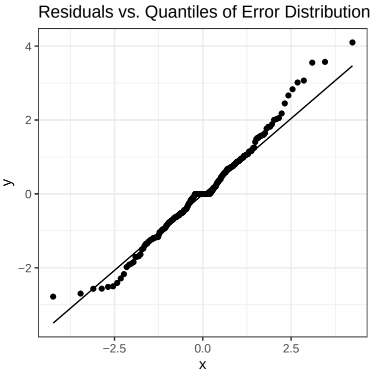
MLE for Laplacian errors = ERM with L1-loss.

- Some losses correspond to more complex or less known error densities, like the Huber loss ▸ Meyer 2021
- Huber density is (unsurprisingly) a hybrid of Gaussian and Laplace

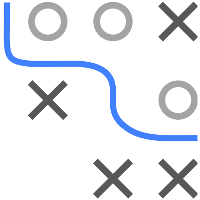
LAPLACE ERRORS - L1-LOSS

- We simulate data $y \mid \mathbf{x} \sim \text{Laplacian}\left(f_{\text{true}}(\mathbf{x}), 1\right)$ with $f_{\text{true}} = 0.2 \cdot \mathbf{x}$.
- We can plot the empirical error distribution, i.e. the distribution of the residuals after fitting a regression model w.r.t. $L1$-loss.
- With the help of a Q-Q-plot we can compare the empirical residuals vs. the theoretical quantiles of a Laplacian distribution.





MAXIMUM LIKELIHOOD IN CLASSIFICATION



Let us assume the outputs y to be Bernoulli-distributed, i.e. $y \mid \mathbf{x} \sim \text{Bern}(\pi_{\text{true}}(\mathbf{x}))$. The negative log likelihood is

$$
\begin{aligned}
-\ell(\boldsymbol{\theta}) &= -\sum_{i=1}^{n} \log p\left(y^{(i)} \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}\right) \\
&= -\sum_{i=1}^{n} \log \left[\pi(\mathbf{x}^{(i)})^{y^{(i)}} \cdot (1 - \pi(\mathbf{x}^{(i)}))^{(1-y^{(i)})}\right] \\
&= \sum_{i=1}^{n} -y^{(i)} \log[\pi(\mathbf{x}^{(i)})] - (1 - y^{(i)}) \log[1 - \pi(\mathbf{x}^{(i)})].
\end{aligned}
$$

This gives rise to the following loss function

$$
L(y, \pi(\mathbf{x})) = -y \log(\pi(\mathbf{x})) - (1 - y) \log(1 - \pi(\mathbf{x})), \quad y \in \{0, 1\}
$$

which we introduced as **Bernoulli** loss.

DISTRIBUTIONS AND LOSSES

For **every** error distribution $\mathbb{P}_\epsilon$ we can derive an equivalent loss function, which leads to the same point estimator for the parameter vector $\boldsymbol{\theta}$ as maximum-likelihood. Formally,
$\hat{\theta} \in \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) \Leftrightarrow \hat{\theta} \in \arg\min_{\boldsymbol{\theta}} -\log(\mathcal{L}(\boldsymbol{\theta}))$.

But: Other way does not always work: We cannot derive a pdf/error distrib. for every loss – the Hinge loss is one prominent example (some prob. interpretation is still possible ▸ Sollich 1999).

When does the reverse direction hold?
If we can write loss as $L(y, f(\mathbf{x})) = L_{\mathbb{P}}(y - f(\mathbf{x})) = L_{\mathbb{P}}(r)$ for $r \in \mathbb{R}$, then minimizing $L_{\mathbb{P}}(y - f(\mathbf{x}))$ is equiv. to maximizing a conditional log-likelihood $\log(p(y - f(\mathbf{x}|\boldsymbol{\theta}))$ if

1. $\log(p(r))$ is affine trafo of $L_{\mathbb{P}}$ (undoing the $\propto$):
   $\log(p(r)) = a - bL_{\mathbb{P}}(r),\ \ a \in \mathbb{R}, b > 0$

2. p is a pdf (non-negative and integrates to one)